肿瘤精准医学知识数据库的设计与构建一

汪凌[1]，陈新[2]二

（1 浙江大学 化学工程与生物工程学院 杭州 310027）

（2 浙江大学 药物生物技术研究所 杭州 310058）

摘要 目的：整合现有前沿的大量而分散的精准医学知识以形成系统完整的知识数据库，为个体组学数据的临床应用提供依据，旨在最终实现基于组学特征的精准用药推荐。方法：采用 MySQL 数据库管理系统构建数据库，从 FDA 伴随诊断、NCCN 指南、My Cancer Genome、GDSC 四大权威医学资源中手动收集精准用药知识，并将原始数据标准化、结构化后以统一的格式存储。结果：成功设计并构建了肿瘤精准医学知识库，目前共收录 1940 条精准用药指导，涵盖了基因突变等 14 种不同类型的组学特征。结论：精准医学知识数据库收录了肿瘤分子组学特征和治疗策略的关联信息，可为临床上个体化治疗方案的制定提供参考依据。数据库的建立为精准医疗临床决策支持系统的开发奠定了基础。

关键词 肿瘤；精准医学；数据库；组学数据

Design and construction of tumor precision medicine knowledge database

WANG Ling[1], CHEN Xin[2]

(1 College of chemical and biological engineering, Zhejiang University Hangzhou 310027, China)

(2 Institute of pharmaceutical biotechnology, Zhejiang University Hangzhou 310058, China)

Abstract Objective: To integrate substantial but scattered state-of-the-art precision medicine knowledge and form a systematic knowledge network, to support clinical application of individual omics data, aiming at precision medication recommendations. Methods: The database was constructed using MySQL. Precision medicine knowledge from FDA companion diagnosis, NCCN guidelines, My Cancer Genome and GDSC was manually collected in a unified format after being standardized and structured. Results: The tumor precision medicine knowledge base (PMKB) was successfully designed and constructed and has already collected 1940 clinical directives, covering 14 kinds of variations. Conclusion: PMKB collects information relating tumor mutations and therapeutic strategies, which can provide personalized treatments of reference. PMKB is also the base of constructing a clinical decision support system of precision medicine.

Key words Tumor; Precision medicine; Database; Omics data

恶性肿瘤已成为目前威胁人类生存健康的最主要因素之一，因癌症死亡的人数占据全球死亡总人数的八分之一[1]。然而通过手术、化疗、放疗等传统癌症疗法治疗时，由于缺少对影像、病理学检查可及范围外的肿瘤生理状态的认识，医生无法预测患者对于特定干预的疗效，无法判断肿瘤的复发和转移，导致疗效欠佳，毒副作用明显，易耐药复发，预后较差[2]。而精准医疗作为新兴的个体化医疗模式，可从组学水平更

一 国家自然科学基金 (31571356)

二 通讯作者，电子邮箱：xinchen@zju.edu.cn

全面地表征个体肿瘤的分子特征，通过包括组学分析、分子检测、分子病理及大数据分析等一系列综合技术手段，帮助临床选择药物响应良好的患者（包括放化疗和靶向药物），指导临床用药的准确性和安全性，提高癌症诊治效果[3]。下一代测序(Next-Generation Sequencing, NGS)技术的发展使多重基因分型和高通量基因组分析变得更为便捷，从而使临床医生能够及时获取治疗相关的分子信息以选择合适的靶向药物。目前，美国食品药品监督管理局(Food and Drug Administration, FDA)已批准如 Extended RAS Panel[4-5]、F1CDx[6-8]等多项基于 NGS 的伴随诊断(Companion diagnostics)检测产品，可通过检测多达数百个的特定基因指导临床精准用药。其他国际权威医学或研究机构也积极推动着精准医疗的临床实践应用，如美国国立综合癌症网络(National Comprehensive Cancer Network, NCCN) 已将精准治疗相关用药策略纳入临床路径指南，My Cancer Genome 整合了精准医学知识以提供肿瘤突变与药物响应性的关系[9-10]，Sanger 研究所与麻省总医院癌症中心合作建立了基于癌症基因组的药物敏感性预测模型(Genomics of Drug Sensitivity in Cancer, GDSC)[11]等。

然而，目前我国精准医疗的临床应用尚未成熟，主要原因是组学数据未得到有效解读和利用，难以与相关精准医学知识进行关联匹配形成明确的治疗策略参考。为解决以上问题，亟待开发一个精准医学知识搜索匹配系统作为临床决策支持(clinical decision support, CDS)工具[12]，以准确联结患者的组学数据与相应的精准用药指导，为个体化治疗方案的制定提供参考依据。但现有的精准医学知识来源分散且在不断更新中，因此建立一个系统整合前沿精准医学知识的数据库成为构建上述搜索系统的必要基础之一。本文将具体阐述肿瘤精准医学知识数据库(Precision Medicine Knowledge Base, PMKB)的数据来源、结构设计及原理、构建方法。PMKB 主要解决了临床用药指征中包含的组学特征数据类型不同、逻辑关系复杂等实际问题，实现了不同数据源的精准医学知识的结构化存储与快捷搜索调用，同时确保数据的完整性与准确性。

1 数据库的设计

1.1 数据来源

目前，本数据库收录了来自四大权威机构的精准医学知识数据资源，分别为 FDA 含伴随诊断的药物标签(label)、NCCN 临床实践指南、My Cancer Genome 精准用药知识以及 Sanger 研究所等提供的 GDSC 精准用药预测资源。

FDA 的伴随诊断是一种与靶向药物相关联的体外诊断技术，主要通过检测人体内蛋白、突变基因的表达水平，在不同类型的疾病人群中筛选出最佳用药人群，有针对性地进行个体化医疗[13-14]。如最早的伴随诊断始于 1988 年 FDA 批准的靶向药物赫赛汀(Herceptin)， 只有通过免疫组化检测确认 HER2 蛋白过表达或通过原位杂交法检测出 HER2 基因扩增的乳腺癌患者才被允许使用赫赛汀治疗。FDA 药物标签全面涵盖了该药的适应证、伴随诊断、用药剂量、注意事项等信息，是肿瘤精准用药指南的权威可靠来源之一。

NCCN 临床实践指南是由 27 个美国知名癌症中心联合制订的癌症临床治疗路径规范[15]，并且指南内容会根据医学进展不断更新以确保其时效性，具有高度临床参考价

值。NCCN 指南覆盖的癌症种类全面，尤其是对于非小细胞肺癌这类靶向药物应用较多的适应证而言，指南中会给出不同分子分型对应的可选治疗方案，并标注其证据等级以区分推荐优先级[16]。

My Cancer Genome 是为医护人员、患者、研究人员提供癌症精准医学知识的一站式工具，主要提供了肿瘤突变与其对应治疗药物的关联信息[9-10]。My Cancer Genome 依据癌种分类，分别给出癌症相关的突变及其亚型、可用药物及其响应性、对应证据、可参与临床试验等信息[17]，可作为 FDA 和 NCCN 两个临床指南级精准用药指导的详细说明与补充资源。

GDSC 资源是英国 Sanger 研究所和美国麻省总医院癌症中心合作建立的基于癌症基因组的药物敏感性数据库，整合收录了大量生物标志物与药物敏感性之间的关系，旨在发掘具有临床意义的治疗标志物用以判断不同患者对治疗的响应性[18]。该项目在超过 1000 个癌细胞系中对 265 种药物进行敏感性测试，并建立了高精度的药物响应预测模型[11]。该资源可为组学特征无明确临床指南匹配(FDA、NCCN)的患者提供细胞系水平的精准用药参考证据。

1.2 数据库结构设计

精准医学知识库的结构设计主要解决以下三个问题：一，PMKB 整合了四大来源的数据资源，如何实现不同来源数据的统一结构化存储，并保证数据的完整性和准确性；二，精准医学知识中的用药指征通常较为复杂，包括多个分子组学特征或其他临床指征，如何准确表征不同类型的组学特征并保存其相互间复杂的逻辑关系；三，如何设计数据表间的关系从而实现数据的快捷搜索调用，即在后续搜索匹配过程中可一次性读取一条用药指导相关的所有用药指征和治疗策略信息。

为实现上述数据结构化存储和快捷搜索调用的目的，本数据库设计了 21 张数据表，其实体关系如图 1。其中，临床用药指导表作为数据库的核心表主要关联了患者基于肿瘤分子水平的用药指征和治疗策略，而注释表存储了每条用药指导的相应文本描述以保证其原始性。

临床用药指导表，存储精准用药指导关联信息，包含 3 个字段：临床用药指导 ID（主键）、综合指征 ID（外键）和治疗策略 ID（表 1）。临床用药指导 ID 同时作为注释表的外键。其他数据表均为此表的扩展表，通过连接查询形成一条完整的临床用药指导。默认情况下所有表均使用自增 ID 作为主键。

由于临床路径或治疗指南中相应的用药指征往往不是单一的，通常包含多个分子病理特征或其他临床指征，因此我们通过综合指征、综合指征成分、分子指征三张数据表将复杂用药指征拆分为多个分子指征并表征其逻辑关系。

综合指征表包含逻辑运算符，该字段值限定范围为 and、or、not、is。

综合指征成分表是综合指征表的扩展表，包含综合指征 ID（外键）、成分类型、组合顺序、分子指征 ID（外键）或综合指征 ID（外键）。其中，“成分类型”字段值限定为综合(complex)或分子(atomic)指征。若成分类型为综合指征，则通过综合指征 ID 外接到综合指征表；若成分类型为分子指征，则通过分子指征 ID 外接到分子指征表。

分子指征表包含特征类型，存储肿瘤分子变异特征的不同类型。特征类型包括高甲基化、拷贝数变异、基因融合、基因表达异常、蛋白表达异常、信号通路激活状态、基因突变、外显子突变、单核苷酸多态性、其他临床指征（如肿瘤分期、治疗史）等 14 种类型。由于不同类型变异所含信息存在较大差异，因此我们创建了 14 张组学特征(feature)表以存储不同特征类型的具体变异信息。每张表包含的字段根据其特征类型专门设计，如外显子突变表包含字段：分子指征表 ID（外键）、基因名 gene symbol、外显子号、突变类型（包括突变 mutation、插入 insertion、缺失 deletion 、跳跃 skipping 等），而信号通路激活状态表包含字段：分子指征表 ID（外键）、通路、状态类型（包括上调、下调）。

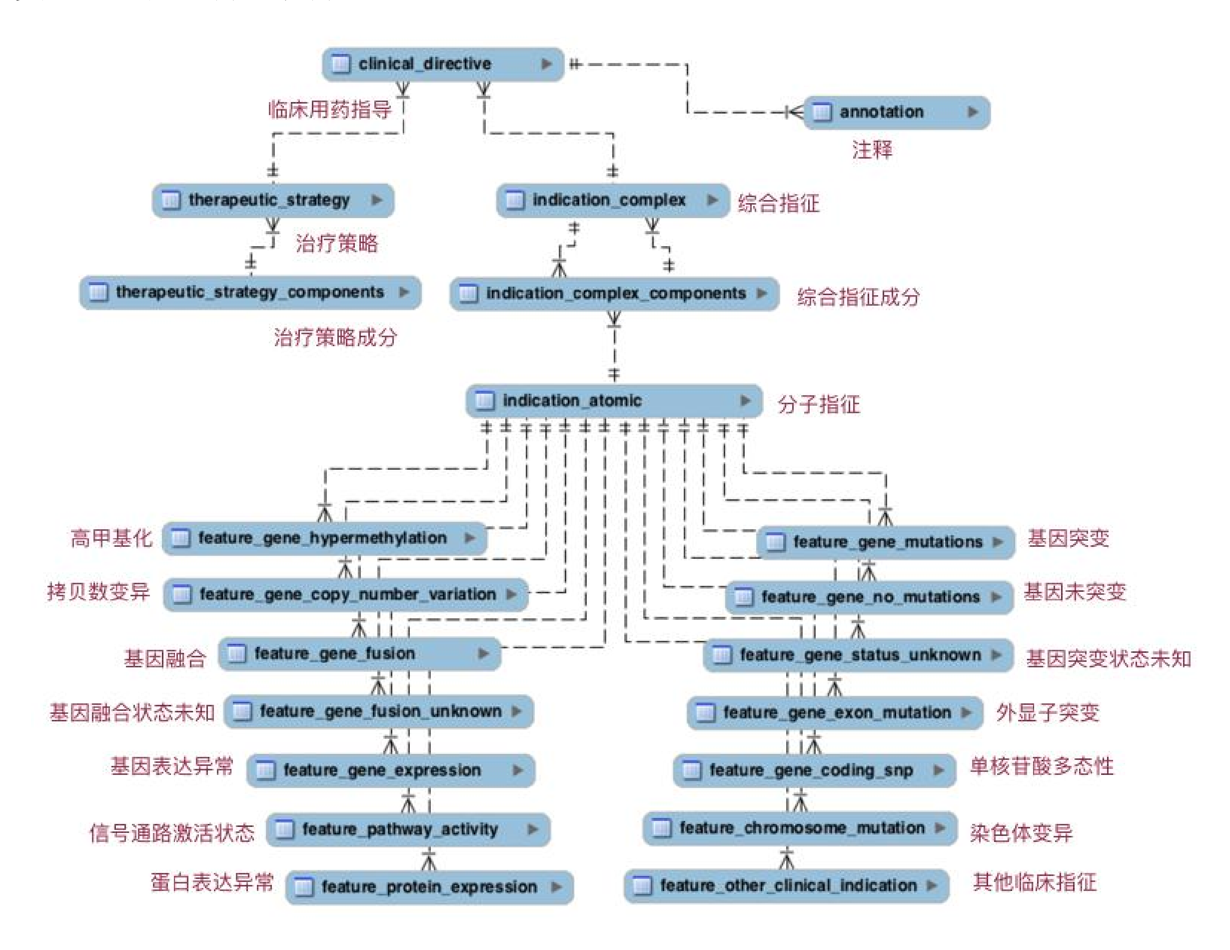


图 1 精准医学知识数据库实体关系图

Fig 1. Entity relationship diagram of PMKB

表 1 临床用药指导表

Table 1 Clinical directive table

临床用药指导 ID Clinical directive ID	综合指征 ID Indication Complex ID	治疗策略 ID Therapeutic Strategy ID
CD1	CI1	TS1
CD2	CI2	TS2

用药指征的复杂逻辑拆分方法具体举例如下：某条临床路径为“当患者肿瘤组织同时发生 A 基因突变和 B 基因突变而无 C 基因突变时，推荐使用 xx 治疗策略”，则此用药指征可结构化为逻辑组合(A and B) or not(C) (图 2)。所有逻辑运算符都被记录在综合指征表中（表 2），逻辑运算符关联的对象（即综合指征或分子指征）被记录在综合指征成分表中（表 3）。逻辑运算符 or 的操作对象是综合指征 CI2 和 CI3，and 的操作对象是分子指征 AI1 和 AI2，not 的操作对象是分子指征 AI3，以上逻辑拆分的顺序记录在综合指征成分表中的“组合顺序”字段。在分子指征表中记录特征类型（表 4），分子指征表的 ID 作为外键与不同组学特征表（如基因突变表）相关联。具体变异的特征信息（如 A 基因突变）记录在组学特征表中。

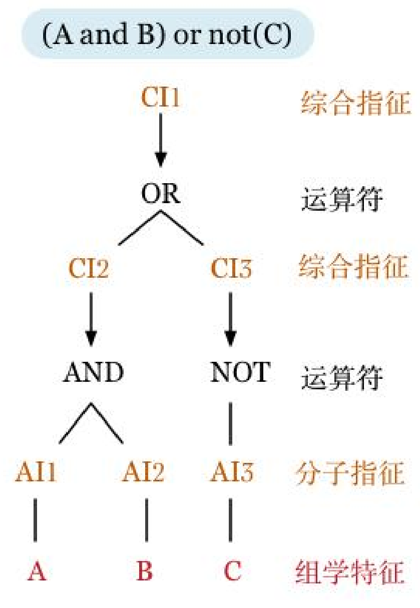


图 2 用药指征逻辑拆分示意图

Fig 2. Logic decomposition illustration of complex indication

表 2 综合指征表

Table 2 Indication complex table

综合指征 ID Indication Complex ID	逻辑运算符 Operator
CI1	or
CI2	and
CI3	not

表 3 综合指征成分表

Table 3 Complex indication components table

综合指征 ID Indication Complex ID	成分类型 Indication Type	组合顺序 Component Order	综合指征 ID Indication Complex ID	分子指征 ID Indication Atomic ID
CI1	complex	1	CI2	
CI1	complex	2	CI3	
CI2	atomic	1		AI1

CI2	atomic	2	AI2
CI3	atomic	1	AI3

表 4 分子指征表

Table 4 Indication atomic table

分子指征 ID Indication Atomic ID	特征类型 Indication Atomic Type
AI1	基因突变
AI2	基因突变
AI3	基因突变

治疗策略表包含治疗策略 ID、治疗策略成分 ID 两个字段（表 5），后者作为外键可外接到其扩展表治疗策略成分表。治疗策略成分表存储具体药物或治疗方法信息，包括治疗策略类型和治疗策略内容（表 6）。其中，治疗策略类型包括化疗、靶向治疗、免疫治疗等多种常见疗法。

治疗策略通常包含不止一种药物，故其所涉及到的药物清单将被展示，目前没有进行逻辑拆分。事实上，如果治疗策略中包含多种药物，可能意味着是药物 A 和 B 联合用药(即 Drug A and Drug B)，也可能是可以使用药物 A 或 B(即 Drug A or Drug B)。为了最大化数据准确性，存在逻辑关系的原始治疗策略信息被储存在注释中以备查询。

表 5 治疗策略表

Table 5 Therapeutic strategy table

治疗策略 ID Therapeutic Strategy ID	治疗策略成分 ID Therapeutic Strategy Components ID
TS1	TSC1
TS1	TSC2

表 6 治疗策略成分表

Table 6 Therapeutic strategy Components table

治疗策略成分 ID Therapeutic Strategy Components ID	治疗策略成分类型 Components Type	治疗策略内容 Therapeutic Strategy Components
TSC1	靶向治疗	Drug A
TSC2	化疗	Drug B

2 数据库的构建方法

本数据库采用 MySQL 数据库管理系统构建。相比于大中型数据库 SQL server 和 Oracle，MySQL 具有功能丰富、使用简便、运行速度快、安全可靠等优势。

由于各来源的精准医学知识大多使用非结构化的自然语言描述，难以准确有效地进行自动化知识抽提与转换， 因此 PMKB 采用手动方式进行数据采集以保证数据的真实有效性。人工收集的 FDA 伴随诊断、NCCN 指南、My Cancer Genome 资源、GDSC

精准用药预测资源，通过数据库开发工具 Navicat 手动加入 PMKB，同时建立各数据表之间的外键联系，从而完成精准医学知识数据的结构化、标准化存储。

3 结果与讨论

精准医学知识库整合了 FDA、NCCN 、My Cancer Genome 和 GDSC 四大权威精准用药资源，并以标准化、可计算的结构存储，以实现肿瘤分子病理特征和治疗策略信息的关联。目前，PMKB 共收录了 1940 条临床用药指导（表 7）、21 张数据表（表 8），涵盖了临床信息、高甲基化、拷贝数变异、基因融合、基因表达异常、蛋白表达异常、信号通路激活状态、基因突变、外显子突变、单核苷酸多态性、染色体变异等 14 种不同类型的分子组学特征。

由于 PMKB 的结构设计具有广泛的通用性与可扩展性，可使不同来源的医学知识数据以结构化的方式统一、 完整 、准确地存储于数据库中。其中，综合指征表、综合指征成分表、分子指征表的设计可有效表征用药指征中的复杂逻辑关系，便于进一步实现患者的真实肿瘤分子组学特征与 PMKB 精准用药知识之间的快速匹配；各组学特征表的字段设计可使不同变异类型的数据在结构化存储（抽提、编码）过程中最大化保留其原始性；各数据表之间的外键关联设计可实现数据的快捷搜索调用，即在后续搜索匹配过程中可一次性读取一条用药指导相关的所有用药指征和治疗策略信息。

表 7 PMKB 临床用药指导条目统计

Table 7 The number of clinical directives collected in PMKB

数据来源	临床用药指导记录条数
FDA	44
NCCN	70
My Cancer Genome	58
GDSC	1768
总计	1940

表 8 精准医学知识数据库条目统计

Table 8 The number of records collected in PMKB

数据表名称	英文表名	记录条数
临床用药指导表	clinical_directive	1940
注释表	annotation	65601
治疗策略表	therapeutic_strategy	499
治疗策略成分表	therapeutic_strategy_components	351
综合指征表	indication_complex	2835
综合指征成分表	indication_complex_components	6006
分子指征表	indication_atomic	2301
高甲基化表	feature_gene_hypermethylation	501
拷贝数变异表	feature_gene_copy_number_variation	359
基因融合表	feature_gene_fusion	12

基因融合状态未知表	feature_gene_fusion_unknown	1
基因表达异常表	feature_gene_expression	1
信号通路激活状态表	feature_pathway_activity	22
蛋白表达异常表	feature_protein_expression	24
基因突变表	feature_gene_mutations	995
基因未突变表	feature_gene_no_mutations	5
基因突变状态未知表	feature_gene_status_unknown	1
外显子突变表	feature_gene_exon_mutation	14
单核苷酸多态性表	feature_gene_coding_snp	19
染色体变异表	feature_chromosome_mutation	1
其他临床指征表	feature_other_clinical_indication	94

我们设计构建的精准医学知识库从海量数据资源中整合、提炼了有效的精准用药知识，可通过匹配肿瘤患者的分子组学数据，为其提供个体化的治疗方案参考。为构建一个完整的临床决策支持系统，我们已建立了病例组学数据库用以存储患者真实的组学数据，并开发了 PMKB 相应的匹配算法以关联患者组学数据与精准用药指导。病例组学数据库、精准医学知识库、匹配算法三者共同构成精准医学知识搜索系统，旨在最终实现临床诊治过程中的精准用药推荐（图 3）。利用此搜索系统，我们已完成 20 例胃癌患者的分子病理分析与精准用药推荐，并将在实际应用中持续优化参数、添加最新的精准医学知识数据。目前 PMKB 采用的手动数据采集方式虽然可以保证数据的准确性，但效率较低，因此我们将在后续工作中利用自然语言处理(natural language processing, NLP) [19]、数据挖掘(data mining) [20]等技术建立自动化的数据采集方法，以实现高效的精准医学知识识别、抽提、编码存储。

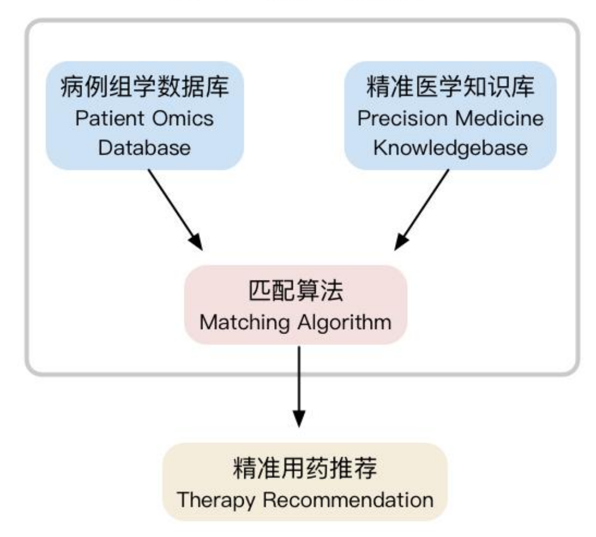


图 3 精准医学知识搜索系统示意图

Fig 3. Illustration of precision medication searching system

参考文献

[1] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*, 2009, 458:719-724.

[2] 付文华, 钱海利, 詹启敏. 中国精准医学发展的需求和任务. 中国生化药物杂志, 2016, 36(4):1-4.

[3] Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. *Journal of Clinical Oncology*, 2013, 31(15):1803-1805.

[4] Sorich MJ, Wiese MD, Rowland A, et al. Extended RAS mutations and anti-EGFR monoclonal antibody survival benefit in metastatic colorectal cancer: a meta-analysis of randomized, controlled trials. *Annals of Oncology*, 2015, 26(1):13-21.

[5] Allegra CJ, Rumble RB, Hamilton SR, et al. Extended RAS gene mutation testing in metastatic colorectal carcinoma to predict response to anti-Epidermal Growth Factor Receptor monoclonal antibody therapy: American Society of Clinical Oncology Provisional Clinical Opinion Update 2015. *Journal of Clinical Oncology*, 2016, 34(2):179-185.

[6] Ali SM, Hensing T, Schrock AB, et al. Comprehensive genomic profiling identifies a subset of crizotinib-responsive ALK-rearranged non-small cell lung cancer not detected by fluorescence in situ hybridization. *Oncologist*, 2016, 21(6):762–770.

[7] Rankin A, Klempner SJ, Erlich R, et al. Broad Detection of Alterations Predicted to Confer Lack of Benefit From EGFR Antibodies or Sensitivity to Targeted Therapy in Advanced Colorectal Cancer. *Oncologist*, 2016, 21(11):1306-1314.

[8] Boussemart et al. Hybrid-capture based genomic profiling identifies BRAF V600 and non-V600 alterations in melanoma samples negative by prior testing. *Annals of Oncology*, 2017, 28(suppl_5):v428-v448.

[9] Kusnoor SV, Koonce TY, Levy MA, et al. My Cancer Genome: Evaluating an Educational Model to Introduce Patients and Caregivers to Precision Medicine Information. *AMIA Joint Summits on Translational Science*, 2016:112.

[10] Levy M, Lovly C, Horn L, Naser R, Pao W. My Cancer Genome: Web-based clinical decision support for genome-directed lung cancer treatment. *Journal of Clinical Oncology*, 2011, 29:15_suppl, 7576.

[11] Iorio F, Knijnenburg T, Vis D, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 2016, 166(3):740.

[12] Castaneda C, Nalley K, Mannion C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, 2015, 5(1):4.

[13] Fridlyand J, Simon RM, Walrath JC, et al. Considerations for the successful co-development of targeted cancer therapies and companion diagnostics. *Nature Reviews Drug Discovery*, 2013, 12(10):743.

[14] Mansfield EA. FDA perspective on companion diagnostics: an evolving paradigm. *Clinical Cancer Research*, 2014, 20(6):1453-1457.

[15] National Comprehensive Cancer Network. About NCCN. 2017. https://www.nccn.org/about/default.aspx

[16] National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology: non-small cell lung. 2017. https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf

[17] My Cancer Genome. 2017. https://www.mycancergenome.org/

[18] Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 2013, 41:D955-D961.

[19] Al-Haddad MA, Friedlin J, Kesterson J, et al. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *HPB*, 2010, 12(10):688-695.

[20] Koh HC, Tan G. Data mining applications in healthcare. *Journal of Healthcare Information Management*, 2005, 19(2):64-72.